D224–D228 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 30 October 2008*
doi:10.1093/nar/gkn833

Protein segment finder: an online search engine for segment motifs in the PDB

Abraham O. Samson* and Michael Levitt

Department of Structural Biology, Stanford University, Stanford, CA 94305, USA

Received August 15, 2008; Revised October 10, 2008; Accepted October 14, 2008

ABSTRACT

Finding related conformations in the Protein Data Bank (PDB) is essential in many areas of bioscience. To assist this task, we designed a search engine that uses a compact database to quickly identify protein segments obeying a set of primary, secondary and tertiary structure constraints. The database contains information such as amino acid sequence, secondary structure, disulfide bonds, hydrogen bonds and atoms in contact as calculated from all protein structures in the PDB. The search engine parses the database and returns hits that match the queried parameters. The conformation search engine, which is notable for its high speed and interactive feedback, is expected to assist scientists in discovering conformation homologs and predicting protein structure. The engine is publicly available at http://ari.stanford.edu/psf and it will also be used in-house in an automatic mode aimed at discovering new protein motifs.

INTRODUCTION

Over the past years, structural data in the Protein Data Bank (PDB) has grown enormously (1). This is due both to the worldwide structural genomics effort, and to recent advances in X-ray crystallography, such as high-intensity synchrotron beam lines. At the present time, approximately 52 000 protein and nucleic acid structures are available (release of August 2008). So much data makes it difficult to navigate the information and realize its wealth, particularly when one tries to retrieve segment motifs (2). Retrieval of such segment motifs is often used in widespread applications such as protein structure prediction, loop modeling and homology matching.

To assist this task and sort quickly and efficiently through available structures, computational engines providing fast structural queries of the PDB have been designed. One such basic engine is the advanced search option of the PDB which handles queries for amino acid sequence and secondary structure amount (3). A more advanced program is MSDMOTIF which can combine searches for sequence motifs, structure motifs, protein sequence, 3D properties secondary structure elements, etc. (4). Although very reliable, the program is resource-intensive and response is very slow limiting its use. Another advanced program is SPASM which was developed by Kleywegt and coworkers (5) to find spatial motifs consisting of arbitrary main-chain and side-chains conformation in a database of protein structures. This program is fast and excels in finding spatial homologs displaying low RMS deviations for a set of PDB coordinates. An additional search engine named Fragment Finder was designed by Sekar and coworkers (6) to identify similar 3D structural motifs. This program is based on the similarity of backbone φ and ψ dihedral-angles and allows superimposed display of search results. Another engine, PAST, which was developed by Griebsch and coworkers (7), is based on translation- and rotation-invariant representation of protein backbone. Takahashi and coworkers (8) developed a 3D substructure search program named SS3D-P2 to find protein motifs based on secondary structure elements. Akutsu *et al.* (9) developed another engine to rapidly search for protein segment homology based on Fourier transforms. In this engine, the similarity of segments is evaluated from the difference between hash vectors consisting of low frequency components of Fourier-like spectrum for the distances between the Cα atom and the centroid. Last but not least, Helmer-Citterich *et al.* designed a search engine named PdbFun for structural and functional analysis of proteins at the residue level. PdbFun executes searches using various criteria such as secondary structure, residue type, protein domain, solvent exposure, ligand binding ability and catalytic activity (10). The aforementioned engines were all designed to identify either spatial similarity or sequence similarity but do not allow combined search for segment motifs using primary, secondary and tertiary structure constraints. There is a clear and so far unmet need for a fast search engine, which combines these query constraints and includes amino acid sequence, secondary structure and contacts.

Here, we describe an online search engine that rapidly finds peptide segments which satisfy a set of

*To whom correspondence should be addressed. Tel: +1 650 725 0754; Fax: +1 650 723 8464; Email: avraham.samson@stanford.edu

conformational parameters in available structural data of the PDB. Query parameters include amino acid sequence, sequence motifs, secondary structure, secondary structure elements, disulfide bonds, hydrogen bonds and residue contacts. Public access to the search engine is facilitated through a simple interactive graphic user interface (GUI). The search engine, named Protein Segment Finder (PSF), is advantageous due to its speed, generality and simplicity. It is expected to be helpful to the scientific community by easing the identification of segment motifs (2) and conformation homologs, and distilling useful information from the PDB.

DATABASE

To avoid the time-consuming task of calculating protein contacts for every search, a database containing contact maps of all PDB proteins was prepared. The database contains entries for all protein structures in the PDB, in the format shown in Figure 1. Each database entry is headed by a line consisting of the PDB ID and the protein name separated by a vertical bar. This header is followed by two lines, one with the amino acid sequence and another with the secondary structure both in the one letter code. The amino acid sequence is all in capital letters, except for cystine pairs which are denoted by small letter pairs. The secondary structure sequence, corresponding to the sequence of amino acids with a particular secondary structure, was calculated using the DSSP program by Kabsch and Sander (11) in which H represents α-helix, E extended β-structure, B isolated β-structure, T hydrogen bonded turn, S bend, I π-helix, G 3_{10}-helix and spaces random coil.

Next comes the contact map which is currently limited to a size of 30 lines for reasons of speed. This contact map summarizes all protein contacts separated by up to 30 residues. Each line of the contact map begins with a dot followed by an index number composed of two digits. The two digit index number corresponds to the number of amino acids separating all interactions on that line. Contact types are recorded using a single code digit ranging from '1' to '8' that denotes the occurrence of heavy atoms and Cα interactions as well as hydrogen bonds. These code digits are defined as follows: '1' indicates contact between heavy atoms; '2' indicates a contact between Cα pairs; '3' indicates a backbone HN···CO hydrogen bond; '4' indicates a backbone CO···NH hydrogen bond; '5' indicates the presence of both these backbone hydrogen bonds. '6', '7' and '8' indicate interactions similar to '3', '4' and '5' except they also indicate the presence of a heavy atom interaction. The PSF database is based on the PDB version of August 2008. It is expected, that this database will be manually updated every 6 months, until an automatic update is programmed. The database was prepared using Perl scripts and the C++ DSSP script (11).

SEARCH ENGINE ALGORITHM

To extract segments obeying the query parameters from the aforementioned database, a search engine was designed using Perl. The engine parses over the entire database and searches for matches of amino acid

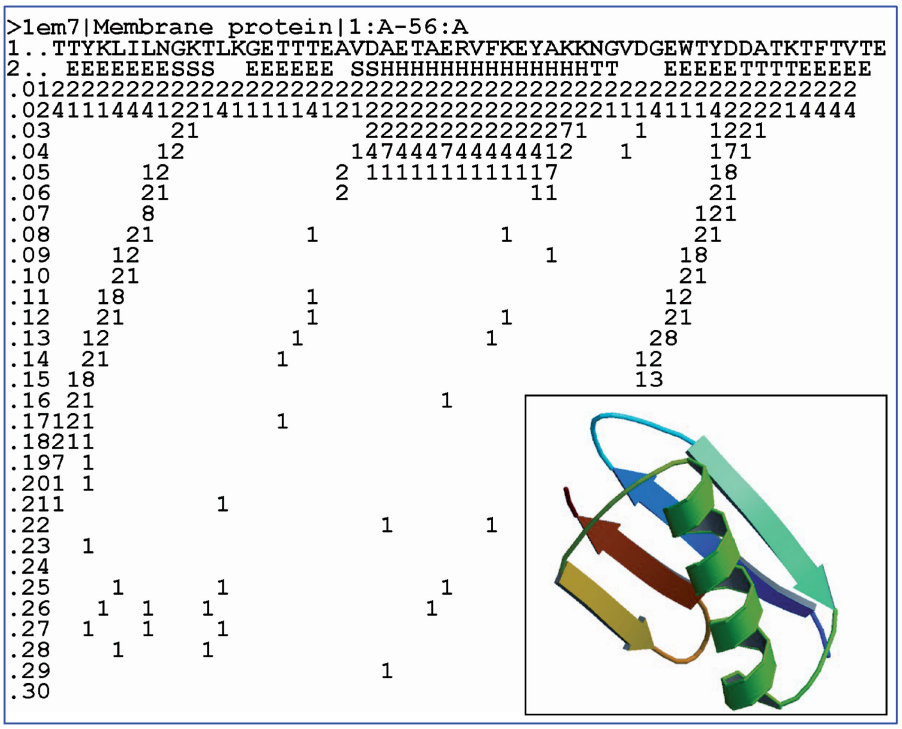

Figure 1. Sample database entry and structure for a typical small protein. Shown is the database entry for PDB ID 1em7 together with the ribbon diagram of the same protein (in inset). Line 1 is the sequence, line 2 is the secondary structure and lines 3–32 indicate contacts between position 1 and a position up to 30 residues further along the chain.

sequence, secondary structure and contacts in a procedural manner. First, the program attempts to match the amino acid sequence. If successful, the program proceeds and attempts to find a secondary structure match, else the next database entry is read. If also the secondary structure is matched then the program continues to find contact matches, else the next database entry is read. This cycle is repeated until all database entries are read. All matches of sequence, secondary structure and contact are stored in three separate lists. These three lists are then compared for common sequence position of the matched segments. If a sequence position is identical in all three lists, then the PDB ID and sequence position of the match is stored in a final hit list. This hit list information is then forwarded to the output manager which in turn displays it textually and graphically.

GRAPHIC USER INTERFACE

To facilitate public use of the program, a GUI was designed using Javascript. In this GUI, the user is prompted to select a query type, from the following: segment size, amino acid sequence, secondary structure sequence, secondary structure element and contacts (Figure 2). Querying for segment size prompts the user to select a length between 1 and 30 residues. Querying for an amino acid or a secondary structure sequence prompts the user to enter the sequence in the one letter code. Querying for a secondary structure element prompts the user to choose among predefined ones such as α-helix, α-hairpin, β-strand, β-hairpin, random coil, turn, π and 3_{10}-helices.

Selecting a query for contacts launches a simplistic molecular display of the queried segment in which contacts are added by mouse clicking on interacting atom pairs (Figure 2). Up to four types of contacts may be added in this manner, namely heavy atom contacts by clicking on the R groups of interacting residues (Figure 2), Cα–Cα interactions by clicking on Cα pairs, backbone hydrogen bonds by clicking on HN and CO atoms, and disulfide bridges by clicking on sulfur atom pairs. Note that because residue contacts are defined as any pair of heavy atoms in the two residues being closer than 6 Å, the atom marker 'R' is essentially any heavy atom in the residue. Upon clicking on the atom pairs a burgundy, red, blue or yellow line connecting the interacting atoms will appear representing Cα, heavy atoms, hydrogen bonds and disulfide contacts, respectively.

For each query type an estimation of the number of matches is available for preview by clicking on the 'evaluate subquery' button. The 'evaluate sub-query' is not a prerequisite for job submission, but rather a means to estimate the number of segment that match the given constraints. An estimate of the total number of hits is obtainable by clicking on the 'evaluate' button. To include more or less query types, click the '+' and '−' button, respectively. For expediency, example queries have been prepared and are available by clicking the 'example' buttons. A tutorial and help messages are available online.

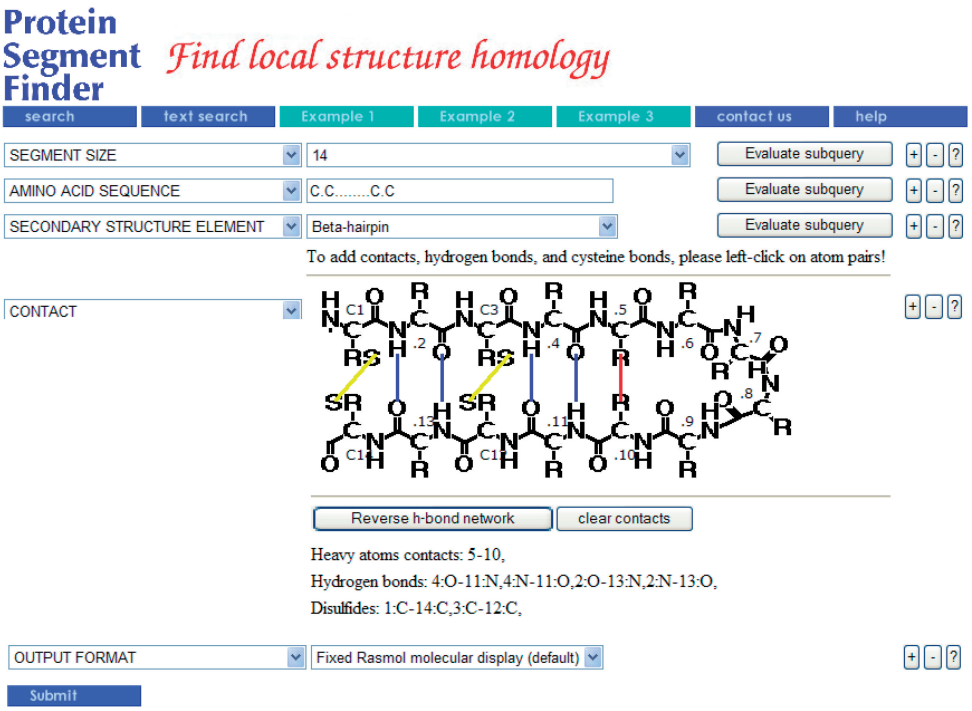

Figure 2. Representative query example. Shown is the query form for a segment of 12 residues that adopts a β-hairpin conformation with the sequence **CxCxxxxxCxC** ('x' represents any amino acid), two disulfide bonds (in yellow), four hydrogen bonds (in blue) and one heavy atom interaction (in red) as drawn.

DATA RETRIEVAL

After entering all the information, the user can run the query by clicking the 'submit' button. The query parameters are then passed on to the search engine using CGI and the run is initiated. Upon run completion, query matches are displayed in text and figures (Figure 3). Text output includes primary and secondary sequence, as well as the relevant contact map. Graphic output includes a Rasmol (12) generated image or a Jmol interactive molecular viewer applet allowing easy viewing of the matched segments. To allow handling of the voluminous results, an adjustable paging system is enabled. The GUI does not require any prior knowledge of scripting languages and permits public and general use of the engine. An alternative to the GUI is a nongraphic text interface, which is available by clicking on the 'text only' button in the output format. This featured interface offers an easy solution for browsers not supporting Javascript. Finally, the database content and its contact map may be retrieved by entering the PDB ID in the query type option 'retrieve PDB entry'.

SERVER CONFIGURATION

The search engine described above is publicly available at http://ari.stanford.edu/psf as a community service. The engine runs on a small server consisting of a Linux operated desktop computer with a 1.8 GHz Intel Pentium processor, with 2 GB of RAM. The web interface is powered by Apache HTTP server version 2.2. Typical search durations are <20 s.

PREDEFINED SECONDARY STRUCTURE ELEMENTS

As mentioned earlier, it is possible to select predefined secondary structure elements, such as β-hairpin, β-strand, α-hairpin, α, π and 3_{10}-helices. These predefined

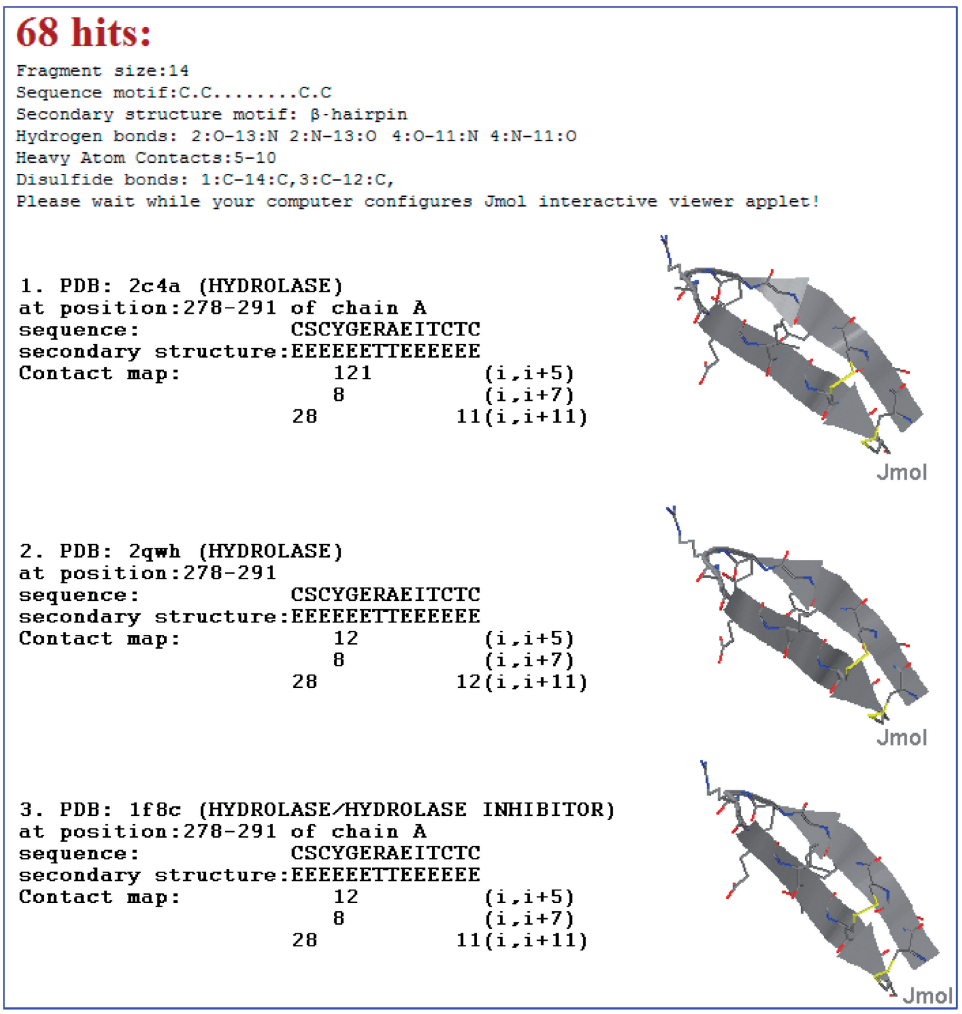

Figure 3. Representative output example. Shown is the output of the query example (Figure 2) for the amino acid sequence **CxCxxxxxCxC** in a β-hairpin conformation. The additional query parameters (two disulfide bridges, four hydrogen bonds and one heavy atom contact) are summarized in the header of the output. Only the three first hits of a total of 68 are shown. For each hit, the output display includes the PDB ID, the sequence position and chain, the amino acid sequence, the secondary structure, the relevant contact map and a Jmol interactive molecular viewer applet for viewing the segment structure.

elements are simply defined as multiple repeats of the basic secondary structure unit, with the exception of α- and β-hairpins. The length of these elements is set by choosing an appropriate 'segment size' between 1 and 30 residues. The program defines β-hairpins as a segment containing a secondary structure of the form EXE in which E is a β-strand and X is a gap, a bend, or a turn, and displaying at least one hydrogen bond at its base. For segments of even length, $n \geq 4$, the β-hairpin secondary structure is defined by the following palindromic regular expression E{1}(E|T|S|\s){$(n-4)/2$}(T|S|\s){2}(E|T|S|\s){$(n-4)/2$}E{1}. For segments with odd number length, $n \geq 3$, the β-hairpin secondary structure is defined by the following palindromic regular expression E{1}(T|S|\s){$(n-3)/2$}(T|S|\s){1}(T|S|\s){$(n-3)/2$}E{1}. To ensure β-hairpin closure and adequate hydrogen bonding, the first and last residues of the hairpin are connected by at least one hydrogen bond. This hydrogen bond may be erased when defining contacts and other hydrogen bonds of choice may be added instead. Notably, this hydrogen bond may be moved by one index, thus inverting the β-hairpin by clicking on the button labeled 'reverse hydrogen-bond network'.

Correspondingly, α-hairpins are defined by the program as a segment containing a secondary structure of the form HXH in which H is an α-helix and X represents a gap, a bend or a turn, with at least one contact between the helices. For segments of even length, $n \geq 4$, the secondary structure is defined as H{1}(H|T|S|\s){$(n-4)/2$}(T|S|\s){2}(H|T|S|\s){$(n-4)/2$}H{1}, and for segments of odd number length, $n \geq 3$, the secondary structure is defined as H{1}(H|T|S|\s){$(n-3)/2$}(T|S|\s){1}(H|T|S|\s){$(n-3)/2$}H{1}. To ensure proper alignment, the first and last residues of the α-hairpin have at least one contact. This contact may be erased when defining contacts. Among the strengths of the search engine is the ability to identify α- and β-hairpins in a rapid and effective manner.

CONCLUSIONS

Over the past decade, we have witnessed the development of methods for fold classification such as SCOP (13) and CATH (14). More recently, however, attention has shifted to structural similarities at an atomic level rather than of the domain fold, such as the conformation of protein segments. Whereas the overall fold is indisputably significant as a framework upon which protein function lays, the actual protein function is usually carried out by a relatively small amount of residues or a protein segment. The search engine presented herein allows the easy and quick identification of such conformational segments. It is expected to be beneficial to the scientific community for comparative structural analysis, for the analysis of protein segments, and for the prediction of structure and function of uncharacterized proteins. The search engine is particularly valuable for finding homologs of NMR structures, which are based on a large number of contacts. In the future, we anticipate a publicly available PSF program for download and use on the client side. We also intend to improve the search engine by allowing search for RNA motifs as well as larger protein segments, thus enabling a more comprehensive survey.

ACKNOWLEDGEMENT

We thank Prof. Jacob Anglister and Dr Osnat Rosen for testing PSF.

FUNDING

National Institutes of Health (GM63817 and EY016525). Funding for open access charge: National Institutes of Health (GM63817).

Conflict of interest statement. None declared.

REFERENCES

1. Levitt,M. (2007) Growth of novel protein structural data. *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.
2. Levitt,M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, **226**, 507–533.
3. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.
4. Golovin,A. and Henrick,K. (2008) MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, **9**, 312.
5. Kleywegt,G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, **285**, 1887–1897.
6. Ananthalakshmi,P., Kumar Ch,K., Jeyasimhan,M., Sumathi,K. and Sekar,K. (2005) Fragment Finder: a web-based software to identify similar three-dimensional structural motif. *Nucleic Acids Res.*, **33**, W85–W88.
7. Taubig,H., Buchner,A. and Griebsch,J. (2006) PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.*, **34**, W20–W23.
8. Kato,H. and Takahashi,Y. (1997) SS3D-P2: a three dimensional substructure search program for protein motifs based on secondary structure elements. *Comput. Appl. Biosci.*, **13**, 593–600.
9. Akutsu,T., Onizuka,K. and Ishikawa,M. (1997) Rapid protein fragment search using hash functions based on the Fourier transform. *Comput. Appl. Biosci.*, **13**, 357–364.
10. Ausiello,G., Zanzoni,A., Peluso,D., Via,A. and Helmer-Citterich,M. (2005) pdbFun: mass selection and fast comparison of annotated PDB residues. *Nucleic Acids Res.*, **33**, W133–W137.
11. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
12. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
13. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
14. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.